DOCUMENT RESUME

ED 364 608 TM 020 880

AUTHOR Huston, Holly L.
TITLE Meaningfulness, Statistical Significance, Effect Size, and Power Analysis: A General Discussion with Implications for MANOVA.
PUB DATE Nov 93
NOTE 29p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (22nd, New Orleans, LA, November 9-12, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Effect Size; *Estimation (Mathematics); *Multivariate Analysis; *Research Methodology; Research Reports; *Statistical Significance
IDENTIFIERS *Meaningfulness; *Power (Statistics)

ABSTRACT

This paper begins with a general discussion of statistical significance, effect size, and power analysis; and concludes by extending the discussion to the multivariate case (MANOVA). Historically, traditional statistical significance testing has guided researchers' thinking about the meaningfulness of their data. The use of significance testing alone in making these decisions has proved problematic. It is likely that less reliance on statistical significance testing, and an increased use of power analysis and effect size estimates in combination could contribute to an overall improvement in the quality of new research produced. The more informed researchers are about the benefits and limitations of statistical significance, effect size, and power analysis, the more likely it is that they will be able to make more sophisticated and useful interpretations about the meaningfulness of research results. One table illustrates the discussion. (Contains 37 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 364 608

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Holly L. Huston

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Meaningfulness, Statistical Significance, Effect Size, and Power Analysis: A General Discussion with Implications for MANOVA

Holly L. Huston

Texas A&M University

Paper presented at the Mid-South Educational Research Association conference, November 9-12, 1993, New Orleans, La.

Running Head: SIGNIFICANCE, EFFECT SIZE, AND POWER ANALYSIS

TM 020 880

BEST COPY AVAILABLE

Meaningfulness, Statistical Significance, Effect Size, And Power Analysis: A General Discussion with Implications for MANOVA

Significance testing, the golden calf of social science statistics, has come under increasingly heavy scrutiny in the last few decades. Criticism of traditional significance testing and its (some say) excessive role of importance in social science research and publication has been both severe and relentless (Carver, 1978; Keaster, 1988; etc). Many critics have argued that too many researchers (Thompson, 1989; Wilkinson, 1992), editors (Kupfersmid, 1988; Sedlmeier & Gigerenzer, 1989), and graduate student dissertation committees (Shaver, 1980; Thompson, 1988) confuse statistical significance of results with practical significance, thereby failing to think more carefully about their data and its relevance to the world at large.

Some critics of significance testing have advanced other statistical methods of aiding researchers in determining the uniqueness and importance of their findings. These other methods have included power analysis and magnitude of effect measures (measures of effect size), two related techniques which are discussed in this paper. Each method has its advantages and limitations, as well as its supporters and detractors. This paper begins with a general discussion of statistical significance, effect size, and power analysis and concludes by extending the discussion to the multivariate case (MANOVA).

The Problem with Significance Testing

A useful conceptual description of the limited role of statistical significance testing was offered by Carver (1978): "Statistical significance simply means statistical rareness. Results are 'significant' from a statistical point of view because they occur very rarely in random sampling under the conditions of the null hypothesis...by itself, statistical significance means little or nothing" (p. 383).

Critics point out, however that statistical significance has become confused with a variety of other things including magnitude of a result (Keaster, 1988), substantive meanings of a result (Keaster, 1988), generalizability of results (Thompson, 1989a), and expected replicability of results (Thompson, 1989b). As Keaster (1988) points out, significance testing yields only a binomial "yes/no" decision, and magnitudes of significance cannot be interpreted. Apart from the general misunderstanding about the actual information provided by a significance test (Chow, 1988), significance testing has been subjected to a number of statistical criticisms as well. Schneider and Darcy (1984) list seven factors that influence the outcome of significance tests.

The most influential of the listed factors is sample size. Since tests of statistical significance are so sensitive to this impact, some researchers have argued that results of significance testing should always be reported in the context of sample size (see Thompson, 1989a; Welge-Crow, LeCluyse, & Thompson, 1990)

that is, at what larger sample size would a non-significant result become significant? and vice versa.

Table 1 (Thompson, 1989b) provides an illustration of this important concept. The table presents significance tests associated with varying sample sizes and the same large (33.6%) fixed effect size. The table can be thought of as presenting results for either a multiple regression analysis with two predictor variables (in which case the "r sq" effect size would be called the squared multiple correlation coefficient, R2) or an analysis of variance involving an omnibus test of differences in three means in a one-way design (in which case the "r sq" effect size would be called the correlation ratio, or eta squared).

Insert Table 1 About Here

The table provides results for fixed effect sizes but increasing sample sizes (4, 13, 23, or 33). For the 33.6% effect size reported in Table 1, the result becomes statistically significant when there are somewhere between 13 and 23 subjects in the analysis.

A researcher who does not really understand statistical significance would differentially interpret the effect size of 33.6% when there were 13 versus 23 subjects in the analysis, even though the effect sizes within the table are fixed. Thompson (1989b) points out that "empirical studies of research practice

indicate that superficial understanding of significance testing has actually led to serious distortions, such as researchers interpreting significant results involving large effect sizes!" (p. 2).

Given the supposedly increased level of awareness about the limitations of significance testing, however, Thompson (1988b) cites four practices which fail to demonstrate general acknowledgement of these shortcomings: (1) the counseling and psychological literature reveals a bias against journal articles that fail to report statistically significant results (e.g.,Atkinson, Furlong, & Wampold, 1982); (2) readers of such literature perceive articles citing significant results more favorably (e.g., Nelson, Rosenthal, & Rosnow, 1986); (3) editors of counseling and psychology journals also are much more favorable toward articles that report significant results and (e.g., Kupfersmid, 1988); and as a result of these attitudes and practices, (4) authors refrain even from submitting articles that cannot report statistical significance (e.g., Kupfersmid, 1988).

A spirited defense of significance testing is mounted by Chow (1988) who asserts that tests of significance can be constructively used provided that significance tests are used to make a statistical decision and not a corroborative or explanatory theoretical decision. However, in the process of making his point he implies that there is considerable confusion about what a statistically significant result means.

A proposed alternative to tests of significance in data interpretation, is effect size. Keaster (1988) suggests that effect size always be examined and that "The basic question to be answered when conducting research is to establish 'how much of the dependent variable is accounted for by the independent variable?', or 'what proportion of the variance in the dependent variable is explained by the observed effect'" (p. 8)?

The Promise of Magnitude of Effect Measures

Effect size can be described as an estimate of how much of the dependent variable is accounted for by the independent variable (Davidson & Giroir, 1989), or how much of the dependent variable(s) can be controlled, predicted, or explained by the independent variable(s) (Snyder & Lawson, 1992). Welge-Crow, LeCluyse, and Thompson (1990) state that effect size may be interpreted as an "index of result importance". Effect size measures are also called "measures of strength of association", "measures of strength of relation", and "measures of magnitude of effect (ME)". For purposes of this paper, effect size measures will be referred to as "ME" (magnitude of effect) measures. Mawell, Camp, and Arvey (1981) state that "a primary advantage of measures of strength of association [ME measures] is that they have the potential to reveal whether a statistically significant result reflects a meaningful rather than a trivial experimental effect (p. 525)." Murray and Dosser (1987) explain that magnitude of effect measures were originally developed with and

for ANOVA procedures. Today, there are ME measures to accompany almost every conceivable statistical test (Prentice & Miller, 1992).

Maxwell and Delaney (1990) describe two "families" of ME measures: 1) Effect size, and 2) Strength of Association measures. According to Maxwell and Delaney, Effect Size measures include those measures that involve differences between means such as: mean difference indices, estimated effect parameter indices, and standardized differences between means (e.g. Cohen's d). Strength of Association measures include indices that involve proportions of variance, or how much of the variability in the dependent variable(s) is associated with the variation in the independent variable(s). Magnitude of these measures ranges from 0-1. Examples of strength of association measures include eta squared, partial eta squared, omega squared, epsilon squared, R2, partial R2, the Wherry formula, the Herzberg formula and the Lord formula (see Snyder & Lawson, 1992 for a more specific discussion of these formulas.)

The most commonly used measures of effect size are the standard mean difference of power analysis, d, and the correlation coefficient, r (Cohen, 1990; Prentice & Miller, 1992). According to Maxwell and Delaney's methods of categorization, d is a measure of Effect Size, and r is a measure of Strength of Association.

Strube (1988) further divides Strength of Association Measures into two additional categories: non-directional vs. directional estimates. He argues that directional estimates, such as r2, contain information about the direction of an effect as well as its magnitude, while non-directional estimates, such as eta squared and omega squared, contain information about magnitude only. Strube states that the importance of this distinction is based in the fact that, for non-directional ME estimates, the size of that estimate will in part depend on sample size which makes ME estimates from studies with different sample sizes noncomparable.

Haase, Ellis, and Ladany (1989) offer some methods for evaluating an absolute obtained ME. First, the obtained ME can be compared to previous research by determining how the obtained effect in a study compares to those obtained in conceptually similar studies. Second, conventional definitions of ME such as those proposed by Cohen (1988) can be used as a crude basis for comparison. Third, the researcher may decide a priori on a minimum meaningful effect size.

The advantages of ME measures include the fact that they supply different information than tests of statistical significance. Where tests of statistical significance have often incorrectly been assumed to be tests of magnitude of experimental effects, ME measures actually <u>are</u> measures of magnitude of experimental effects! As Snyder and Lawson point out, "A small p

value does not necessarily imply that the strength of the relation between the independent and dependent variables in a particular study is large" (p. 5). When used correctly, ME measures greatly enrich a researchers ability to meaningfully interpret his or her data. Rosnow and Rosenthal (1989) cite two arguments for the use of ME measures, and the reporting of effect size in the research literature, first, computing population effect sizes guides our judgement about the sample size needed in the next study we might conduct (a la' power analysis); and second, a result that is statistically significant is not necessarily practically significant and vice versa. Prentice and Miller (1992) cite more benefits of ME measures: (1) effect size estimates indicate the degree to which a phenomenon is present in a population on a continuous scale, with zero indicating that the phenomenon is absent; (2) effect sizes come with conventions (see Cohen, 1988) as to what values constitute small, medium, and large effects; (3) they provide an indication of the practical significance of an effect unlike standard tests of significance. (4) they can be used to quantitatively compare the results of two or more studies as in meta-analysis; and (5) they can be used in power analyses to determine how many subjects are needed in a particular study.

Like tests of statistical significance, magnitude of effect measures have their own limitations and set of cautions. O'Grady (1982) thoroughly discusses a number of theoretical,

methodological and psychometric limitations in regards to magnitude of effect measures. For example, methodologically, he cautions researchers that within subjects sampling designs produce larger measures of explained variance than between-subjects sampling designs. As a theoretical caution, he points out that if most behaviors researched in the social sciences are multiply determined "the magnitude of measures of explained variance between the operations of the constructs will be limited in size to the maximum amount of the variance shared by the two constructs" (p. 774).

Murray and Dosser (1987) similarly describe some problems with ME measures. Specifically, they argue that eta squared, a commonly used ME measure, is dependent on both study size (number of treatments) and sample size. More generally, they express concern about the conceptual confusion between a parameter and a statistic in ME measures, stating that population and sample measures of ME have sometimes been used interchangeably. From a statistical perspective, they argue that in most cases sampling distributions for sample measures of ME are not well known, and in many cases are not known at all. They continue to argue that without knowledge of the sampling distribution of these measures it becomes impossible to accurately specify a "large" or a "small" ME, though Cohen (1988) has provided crude estimates. Snyder and Lawson (1992) provide some additional cautions regarding the use of ME measures for the interested researcher.

Effect size measures were initially viewed by some supporters as an answer to the limitations and "evils" of statistical significance testing. In actual use, however, though these measures have advantages, they also proved to have limitations, and, like significance tests, do not relieve researchers of the responsibility for determining the meaningfulness of their data (Murray & Dosser, 1987; O'Grady, 1982). According to Maxwell et al., "Although measures of strength of association provide one useful adjunct to statistical significance tests, researchers should be careful not to become slaves to measures of association any more so than to p levels" (p. 533).

The Role of Power Analysis

Rossi (1990) described the power of a statistical test as "the probability that the test will correctly reject the null hypothesis" (p. 646). Power analysis (as a way of determining the statistical uniqueness of one's results) is a statistical newcomer in comparison to statistical significance testing. The importance of and the technique of power analysis have been brought to popular awareness by Cohen (1962, 1969). In a 1992 article, Cohen describes statistical power analysis as follows:

> Statistical power analysis exploits the relationships among the four variables involved in statistical inference: sample size (N), significance criterion (alpha), population effect size (ES), and statistical power. For any

statistical model, these relationships are such that each is a function of the other three ....For any given statistical test, we can determine power for given alpha, N, and ES. For research planning, however, it is most useful to determine the N necessary to have a specified power for given beta and ES...(p. 156). To reiterate, the power of a given statistical test depends on three factors: (1) the alpha level chosen by the experimenter, (2) sample size, and (3) effect size. Like statistical significance, power is heavily dependent on sample size.

According to Rossi (1988), there are several benefits of power analysis. First, knowledge of the power of a statistical test for a particular study indicates the likelihood of obtaining a statistically significant result. This can help the researcher decide whether or not undertaking the study is worthwhile and may save time, effort, and money. Second, knowledge of power aids in interpretation of null results. For example, if power for a given study was low then it can be suggested that there was not a good chance of rejecting the null a priori, and therefore the failure to reject the null should not immediately rule out the alternative hypothesis (i.e., poor power is correlated with an increased probability of making a Type II error). Finally, Rossi argues that statistical power can provide useful information about whole areas of research. When the average statistical power of a body of literature is low, "the veracity of even

statistically significant results may be questioned, because the probability of rejecting a true null hypothesis may then be only slightly smaller than the probability of rejecting the null hypothesis when the alternative is true" (p. 647). An additional benefit of power analysis may be that it forces one to consider the magnitude of effects for a particular study, given that effect size is an integral part of the determination of power, thus increasing awareness of the importance of effect size estimates (Cohen, 1990). It may also make researchers more aware of the relationship between the four factors that determine statistical power.

A "limitation" of power analysis is that it has not been readily accepted and utilized by the vast majority of researchers, most visibly, published researchers (Rossi, 1990; Sedlemeier & Gigerenzer, 1989). An important point bearing on this is that once a researcher has obtained a statistically significant result, it is not possible to make a Type II error. Given the current bias toward statistically significant results which remains in practice in the vast majority of social science journals, the reporting of power analyses then becomes irrelevant. This situation produces what Rosenthal (1979) referred to as "the file drawer problem" , meaning that "the journals are filled with the 5% of the studies that show Type I errors, while the file drawers back at the lab are filled with 95% of the studies that show nonsignificant (e.g., p>,05 results"

(p. 638). Kupfersmid (1988) expresses the concern that Type I error is more serious than Type II error "because when a Type I error appears in print it often stops researchers from studying and/or reporting nonsignificant results" (p. 637).

Maxwell and Delaney (1990) consider experimenters' failure to take advantage of power analysis techniques as more of a problem than the "sample size/significance" problem. Sedlmeier and Gigerenzer, (1989) in a study of the effect of power analysis on research studies in the Journal of Abnormal and Social Psychology, noted that the emergence and positive reception of power analysis has had "no noticeable effect on actual practice" (p. 313). As an explanation for this, these authors cite the fact that Fisher's theory of null hypothesis testing was historically first, and had gotten institutionally entrenched, making it difficult for researchers to 'make the switch' to power analysis. Another possible reason for this cited by Rossi (1990), is that Cohen's (1988) effect size estimates are arbitrarily based and do not reflect actual population effect size parameters for a given area of research. Existing surveys of effect sizes have tended to support Cohen's estimates, but few of these surveys have been done.

Significance Testing, Effect Size, and Power Analysis Revisited

Haase, Ellis, and Ladany (1989) describe the interpretation of the magnitude of experimental effects as a two stage process. First, assess statistical significance; and, second, evaluate the

absolute magnitude of the effect. They argue that in order to perform a good evaluation of the outcome of an experiment, analysis depends on both correctly interpreted tests of statistical significance and a measure of ME. Thompson (1988) points out that the largest errors are made in data interpretation "when sample size is small and effect sizes are large but are underinterpreted, or when sample sizes are commendably large and are statistically significant but effect sizes are modest and are overinterpreted" (p.8). Haase et al. (1989) also briefly mention the importance of power analysis in interpretation. Ideally, a calculation of statistical power could also aid a priori in developing a research design likely to yield meaningful results, and post hoc, in interpreting the data. Utilizing all of these methods to develop and analyze a particular study would require a researcher to become more sophisticatedly familiar with the strengths and the limitations of his or her data. An advantage of this, is that this increased familiarity with the data may aid in interpreting the true "meaningfulness" of the findings in the context of a particular body of research literature.

Implications for MANOVA

Bray and Maxwell (1985) state "The mathematical basis for applying MANOVA is well known and generally accepted... However...[certain] areas are still in development [such as] issues concerning the overall test, such as choice between test

statistics, power and sample size concerns, and measures of effect size..."(p. 8). Currently, there is little in the educational and psychological research literature that specifically addresses magnitude of effect measures or power analysis in multivariate analysis, though the issue of significance testing is more thoroughly addressed.

Significance testing in MANOVA differs from significance testing in ANOVA in that, for ANOVA, the null hypothesis is that the means of the groups are equal to one another. In MANOVA, the null hypothesis is that the population means of the groups are equal to one another for all variables and we are testing to determine if the group centroids are equal. A centroid is a vector of population means for a given group (Bray & Maxwell, 1985). In MANOVA, an overall, or omnibus, significance test is obtained using one of several available test statistics, as well as univariate significance tests for follow-up univariate analyses. Significance testing in MANOVA is similar to significance testing in ANOVA in that all the previously discussed limitations of significance testing still apply. According to Maxwell (1992): "As in ...ANOVA, a statistically significant result is not necessarily an important result" (p. 164).

Bray and Maxwell (1985) report that a number of multivariate measures of strength of association (ME measures) have been proposed in the literature for most multivariate techniques. For

example, an ME measure in multiple regression analysis would be called the "coefficient of determination" or the "squared multiple correlation coefficient" (R2= SOS explained/ SOS total) (Snyder & Lawson, 1992). More detailed discussion of multivariate ME measures, such as multivariate generalizations of eta squared, omega squared, and other ME measures are available from Cramer and Nicewander (1979); Huberty (1972); Stevens (1972); and Tatsuoka (1973). Snyder and Lawson (1992) state that "researchers asking multivariate questions will need to employ magnitude of effect measures that are consistent with their multivariate view of the research problem" (p. 15).

In MANOVA, the most commonly used multivariate ME measure is D2. D2, also known as the Mahalanobis distance, is a "natural squared generalization of the univariate measure d, where the means have been replaced by the mean vectors and s (standard deviation) has been replaced by its squared multivariate generalization of within variability, the sample covariance matrix S" (Stevens, 1986, p. 178). Another way of explaining it is that D2 is Hotelling's T2 without the sample sizes.

D2 is a two group measure of separation of groups that is independent of sample size. In regard to interpreting D2, Stevens notes that for values of D2 $\geq$ .64 and n $\geq$ 25, power is generally poor and not adequate for the .05 alpha level. For D2 specifically, Stevens provides crude estimates of effect size: .25=small effect, .5=moderate effect, and >1=a large effect. He

also describes a possible confound in interpreting an overall multivariate effect size, in that a single large univariate effect size may produce a large overall effect size. Rosnow and Rosenthal (1989) agree, and go further in stating that the omnibus effect size estimates are of questionable value in answering a specific research question.

Stevens (1980) argues that the use of power analysis is as important in MANOVA as it is in ANOVA, however, it is too infrequently used by researchers. Rossi (1990) described the computation of multivariate power as being relatively inaccessible. This may explain why power analysis is so little used with MANOVA. In support of this position, Bray and Mawell (1985) point out that making an a priori estimate of power when planning a MANOVA study is difficult because the numerous parameters involved are generally unknown, and must be estimated, rendering the power analysis procedure for MANOVA much more complicated than for ANOVA. Maxwell (1990) also notes that it is much more difficult to formulate a useful value for effect size in the multivariate case (MANOVA) than in the univariate case due to the number and complexity of the parameters that must be estimated (not only means and standard deviations, but also within-group correlations among the dependent variables.) Additionally, Cohen has not provided power estimates for MANOVA, as he has done for ANOVA (Maxwell, 1990).

In order to help remedy the information shortage in this regard, Stevens (1980) has generated tables for determining power in both a two-group and a k-group case, however, he notes that the tables for the k-group case leave problematic "gaps". Stevens also makes several points about the relationship between power analysis and MANOVA. First, as the number of dependent variables increases, sample size must be increased to maintain a given level of power for a specified effect size. Specifically, he cautions researchers to limit the number of DV's to ten or fewer unless the sample size is very large. Second, as a general caution he reports that the "power of the multivariate tests with small-to-moderate group and effect sizes is poor as is true in univariate ANOVA (p. 736)." Third, the magnitude of the within-group correlations has an effect on power. Power with high intercorrelations tends to be greater than power with low intercorrelations. Stevens also notes that post hoc estimation of multivariate power, using the values supplied in his table, is much simpler than a priori estimation.

Discussion

In multivariate analysis, as in univariate analysis, the question of whether or not a result has practical significance, cannot be solved by any single statistical formulation alone. It is still incumbent upon the researcher to reflect on the data, the findings, and the context in which the research exists in order to determine the meaningfulness of obtained results.

Historically, traditional statistical significance testing has guided researchers' thinking about the meaningfulness of their data. The use of significance testing alone in making these decisions has proved problematic and remains so.

The relatively recent addition of magnitude of effect measures (effect size) and power analysis (both a priori and post hoc) to the statistical arsenal of researchers, though promising greater clarity for the task of determining meaningfulness of results, has been disappointing in actual application. Wilkinson (1992) reports that the reporting of effect sizes in the educational and psychological research literature is relatively rare. This holds especially true for the reporting of multivariate results, so much so in fact, that Stevens (1992) provides a method for calculating D2 (a multivariate ME measure) from the reported multivariate F allowing a reader to obtain his or her own effect size estimates! Cohen (1992) describes that the rate of reporting of power analyses in research, though clearly beneficial to result interpretation, is disappointingly low. Sedlmeier and Gigerenzer (1989) point out that given the paramount importance placed on significant results, it almost seems as though much research is designed to yield only a small chance of a significant result in the presence of a true effect.

Rosnow and Rosenthal (1989) cite three relevant factors which contribute to the current crisis in the "essence of statistical data analysis". First, a general overreliance on

dichotomous significance testing decisions leads to the impression that a sharp line exists between significant and non-significant differences. Second, many research studies are done in situations of low power, virtually guaranteeing that they will to fail to reach a given p value for a particular effect size. Third, an entrenched and institutionalized habit of defining the results of research solely in terms of significance levels distorts interpretation of data by excluding valuable input from power analyses and ME measures.

It is likely that less reliance on statistical significance testing, and increased utilization of power analysis and effect size estimates in combination could contribute to an overall improvement in the quality of new research produced. There are no easy short-cuts for determining the practical importance of research results. Cohen (1990) reminds us:

> It is on your informed judgement as a scientist that you must rely, and this holds as much for the statistical aspects of the work as it does for all others. This means that your informed judgement governs the setting of the parameters involved in the planning (alpha, beta, population effect size, sample size, confidence interval), and that informed judgement also governs the conclusions you will draw (p. 1310).

The key to this admonition is the word "informed". The more informed researchers, editors, and graduate student dissertation

committees are about the benefits and limitations of statistical significance, effect size, and power analysis, the more likely it is that they will be able to make more sophisticated and useful interpretations about the meaningfulness of their research results.

References

Atkinson, D.R., Furlong, M.J., & Wampold, B.E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? Journal of Counseling Psychology, 29 (2), 189-194.

Bray, J.H., & Maxwell, S.E. (1985). Multivariate Analysis of Variance. Beverly Hills, CA: Sage.

Carver, R.P. (1978). The case against significance testing. Harvard Educational Review, 48 (3), 378-399.

Chow, S.L. (1988). Significance Test or Effect Size? Psychological Bulletin, 103 (1), 105-110.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ.: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

Cramer, E.M., & Nicewander, W.A. (1979). Some symmetric invariant measures of multivariate association. Psychometrika, 44, 43-54.

Davidson, B. & Giroir, M.M., (1989, November). Comments on better uses of and alternatives to significance testing. Paper presented at the annual meeting

of the Mid-South Educational Research Association, Little Rock, AR. (ERIC Document Reproductive Service No. ED 314 450)

Haase, R.F., Ellis, M.V., & Ladany, N. (1989). Multiple criteria for evaluating the magnitude of experimental effects. Journal of Counseling Psychology, 36, 511-516.

Huberty, C.J. (1972). Multivariate indices of strength of association. Multivariate Behavioral Research, 7, 523-526.

Keaster, R. (1988, November). Statistical significance testing: From Routine to ritual. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY. (ERIC Document Reproductive Service No. ED 303 514)

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

Maxwell, S.E. (1992). Recent developments in MANOVA applications. In B. Thompson (Ed.), Advances in social science methodology: Vol.2 (pp. 137-168). Greenwich, CN: JAI Press Inc.

Maxwell, S.E., Camp, C.J., and Arvey, R.D. (1981). Measures of strength of association: A comparative examination. Journal of Applied Psychology, 66 (5), 525-534.

Maxwell, S.E., & Delaney, H.D. (1990). Designing experiments and analyzing data: A model comparison perspective. Belmont, CA: Wadsworth Publishing.

Murray, L.W., & Dosser, D. (1987). How significant is a significant difference? Problems with the measurement of magnitude of effect. Journal of Counseling Psychology, 34 (1), 68-72.

Nelson, N., Rosenthal, R., & Rosnow, R. (1986). Interpretations of signifcance levels and effect sizes by psychological researchers. American Psychologist, 41, 1299-1301.

O'Grady, K.E. (1982). Measures of explained variance: Cautions and limitations. Psychological Bulletin, 92 (3), 766-777.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. Psychological Bulletin, 112 (1), 160-164.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. Psychological Bulletin, 86, 638-641.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Rossi, J.S. (1990). Statistical power of psychological research: What have we gained in 20 years? Journal of Consulting and Clinical Psychology, 58, 646-656.

Schneider, A.L., & Darcy, R.E. (1984). Policy implications of using signficance tests in evaluation research. Evaluation Review, 8 (4), 573-583.

Shaver, J. (1980, April). Readdressing the role of statistical tests of significance. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA. (ERIC Document Reproductive Service No. ED 191 904)

Snyder, P., & Lawson, S. (1992, April). Evaluating statistical significance using corrected and uncorrected magnitude of effect size estimates. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproductive Service No. ED 346 123)

Stevens, J.P. (1972). Global measures of association in multivariate analysis of variance. Multivariate Behavioral Research, 7, 373-378.

Stevens, J.P. (1980). Power of the multivariate analysis of variance tests. Psychological Bulletin, 88, 728-737.

Stevens, J.P. (1992). Applied multivariate statistics for the social sciences (2nd ed.), New Jersey: Lawrence Erlabaum.

Strube, M. J. (1988). Some comments on the use of magnitude-of-effect estimates. Journal of Counseling Psychology, 35 (3), 342-345.

Tatsuoka, M.M. (1973). An examination of the statistical properties of a multivariate measure of strength of relationships. Urbana: University of Illinois. (ERIC Document Reproduction Service No. ED 099 406)

Thompson, B. (1988a, November). Common methodology mistakes in dissertations: Improving dissertation quality. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY. (ERIC Document Reproductive Service No. ED 301 595)

Thompson, B. (1988b). A note about significance testing. Measurement and Evaluation in Counseling and Development, 20 (4), 146-148.

Thompson, B. (1989a). Asking "what if" questions about significance tests. Measurement and Evaluation in Counseling and Development, 22, 66-68.

Thompson, B. (1989b). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-6.

Welge-Crow, P.A., LeCluyse, K.B., & Thompson, B. (1990, June). Looking beyond statistical significance: result importance and result generalizability. Paper presented at the annual meeting of the American Psychological Society, Dallas, TX. (ERIC Document Reproductive Service No. ED 320 965)

Wilkinson, R. (1992, January). Statistical significance testing: Alternatives and considerations. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document Reproductive Service No. ED 347 169)

Table 1

Statistical Significance at Various Sample Sizes
for a Fixed Effect Size (Large Effect Size)

Source	SOS	r sq	df	MS	F calc	F crit	Decision
SOSexp	**337.2**	**0.336**	**2**	**168.600**	0.253	200.0	Not reject
SOSunexp	**665.1**		1	665.100			
SOStot	**1002.3**		3	334.100			
SOSexp	**337.2**	**0.336**	**2**	**168.600**	2.535	4.10	Not reject
SOSunexp	**665.1**		10	66.510			
SOStot	**1002.3**		12	83.525			
SOSexp	**337.2**	**0.336**	**2**	**168.600**	5.070	3.49	Reject
SOSunexp	**665.1**		20	33.255			
SOStot	**1002.3**		22	45.559			
SOSexp	**337.2**	**0.336**	**2**	**168.600**	7.605	3.32	Reject
SOSunexp	**665.1**		30	22.170			
SOStot	**1002.3**		32	31.322			

Note. As sample size increases, "critical F" values become smaller. Also, as sample size increases, so does error df. As sample size increases, mean square error becomes smaller, therefore "F calc" becomes larger. Entries in bold are fixed for the purposes of these analyses. From Thompson (1989b), with permission.